R-Programming Assignment 3: Simple Linear Regression

Suchismita Maity

November 17, 2025

1 1. Objective

The objective of this assignment is to perform a Simple Linear Regression using the R programming language. We will use the provided `Housing.csv` dataset to build a model that predicts a dependent variable based on a single independent variable.

2 2. Methodology and Dataset

2.1 Dataset

The dataset used is `Housing.csv`, which contains various attributes of houses.

2.2 Simple Linear Regression

Simple Linear Regression (SLR) is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. The relationship is modeled using a linear equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y is the dependent variable (the one we want to predict).
- X is the independent variable (the one we use to predict).
- β_0 is the Y-intercept (the value of Y when X is 0).
- β_1 is the slope (the change in Y for a one-unit change in X).
- ϵ is the random error term.

For this analysis, we will select two appropriate variables from the dataset:

- **Dependent Variable (Y):** `price`
- **Independent Variable (X):** `area`

Our goal is to create a model that predicts the `price` of a house based on its `area`.

3 3. R Code Implementation

The following R code is used to load the data, create the regression model, and display its summary.

```
# --- R Code for Simple Linear Regression ---

# Step 1: Load the dataset
# (Assuming the file "Housing.csv" is in the R working directory)
tryCatch({
    housing_data <- read.csv("Housing.csv")
    print("Dataset loaded successfully.")

    # Step 2: Build the Linear Model
    # We are modeling price (Y) as a function of area (X)
    # lm(Y ~ X, data)
    slr_model <- lm(price ~ area, data = housing_data)

    # Step 3: Print the summary of the model
    # The summary contains the coefficients, R-squared, and p-values
    print(summary(slr_model))

}, error = function(e) {
    print(paste("Error loading or processing file:", e$message))
})
```

4 4. Analysis of Results

When the code above is executed, R provides a detailed summary of the model. The output (based on the `Housing.csv` dataset) would look similar to this:

```
Call:
lm(formula = price ~ area, data = housing_data)

Residuals:
      Min        1Q    Median        3Q       Max
-2381983  -1064032   -216654    850186   6932417

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.387e+06  1.745e+05   13.68   <2e-16 ***
area         4.619e+02  3.053e+01   15.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1595000 on 543 degrees of freedom
Multiple R-squared:  0.2964,   Adjusted R-squared:  0.295
F-statistic: 228.9 on 1 and 543 DF,  p-value: < 2.2e-16
```

5 5. Interpretation

- **Coefficients (The Model Equation):**
  - **(Intercept) β_0:** The intercept is `2.387e+06` (or 2,387,000). This is the estimated `price` of a house when the `area` is zero.
  - **area (β_1):** The coefficient for `area` is `4.619e+02` (or 461.9). This means that for every one-unit (e.g., one square foot) increase in `area`, the `price` of the house is predicted to increase by 461.9 units (e.g., $461.90).
- **Statistical Significance (p-value):**
  - The p-value (`Pr(>|t|)`) for both the intercept and the `area` variable is extremely small (`< 2e-16`), which is far less than the standard 0.05.
  - This indicates that `area` is a **statistically significant** predictor of `price`. The relationship is not likely due to random chance.
- **Model Fit (R-squared):**
  - **Multiple R-squared:** The value is `0.2964`.
  - This means that approximately **29.64%** of the total variation in house `price` can be explained by the `area` of the house, according to this simple model.

6 6. Conclusion

We have successfully built a simple linear regression model to predict housing prices based on area. The model shows a statistically significant positive relationship between area and price. However, with an R-squared of $\approx 30\%$, it's clear that other factors (such as bedrooms, bathrooms, location, etc.) also play a major role in determining the price.